ISBN-553-0552-1 No 5
Statistics - June 29 1984

PRIORITY QUEUE

by

B. Natvig

ABSTRACT

The present report represents an entry PRIORITY QUEUE which is to appear in the Encyclopedia of Statistical Sciences, Vol.6 published by Wiley in 1985. It reviews classical priority queueing models and three papers having specific applications in mind.

PRIORITY QUEUE

As a general background regarding ideas and notation we refer the reader to the entry QUEUEING THEORY. The theory for priority queueing systems is the branch of queueing theory dealing with systems where high priority customers get faster through the system at the expense of others. The basic reference seems still to be [4] giving a series of different models, with a variety of priority disciplines, a solid mathematical treatment. In [5] priority queueing systems of special interest in computer applications are considered.

Since the applicability of queueing theory often has been questioned we will here give reviews of three papers having specific applications in mind. We are not claiming that these are either the best or the most general in the area. One should also remember that even a complex queueing system is more sympathetic to mathematical modelling than a series of other real-life systems. Hence, queueing theory has been and will still be an experimental area to develop tools which are useful in other areas of applied probability.

CLASSICAL PRIORITY QUEUEING MODELS

Assume that the customers are divided into p different priority classes each having a priority index $1,2,\ldots,p$. The index 1 corresponds to the highest priority class and at the other end p corresponds to the lowest. For simplicity we assume a single server and an infinite waiting room capacity. Within each class we have a first-in-first-out (FIFO) queueing discipline. Denoting the inter-arrival time distribution and the service time distribution within the i th class by the symbols A_i and B_i respectively,

we can use the notation $A_1/\cdots/A_p/B_1/\cdots/B_p/1$ for this class of models.

However, we have to specify the queueing discipline between customers from different priority classes. Compare the i th and j th class, where $1 \leq i < j \leq p$. In the queue all customers from the i th class are standing in front of all customers from the j th and will hence be served first. If on the other hand a customer from the j th class is being served on the arrival of a customer from the i th class, there are different alternatives. The priority discipline is preemptive* if the current service is immediately interrupted and service is started for the arriving customer. It is nonpreemptive if service is continued to completion, and it is discretionary if the server may use his discretion to decide which of the two former strategies to use in each case. For instance he may use the nonpreemptive discipline if the amount of service received exceeds a certain level [4], or if remaining service time is sufficiently small [3].

AN APPOINTMENT SYSTEM

Consider the GI/M/1 queueing model with infinite waiting room capacity. The customer arriving at $t=0$ will find $k-1$ customers waiting. The latter customers belong to the second priority class, whereas the ones arriving in $[0,\infty)$ belong to the first. The priority discipline is nonpreemptive whereas within each class we have a FIFO queueing discipline.

As a motivation for studying the present model consider the following specialization of the arrival pattern above which is realistic when, for example, doctors, dentists or lawyers are consulted. Let the intervals between possible arrivals have fixed

length $1/\lambda$ and let the probability of a customer not turning up be $1-p$. Customers turn up or not independently of each other. The number of intervals of length $1/\lambda$ between the arrivals of two customers are then geometrically distributed with parameter p. The $k-1$ customers in the second priority class do not have an appointment, but are allowed to queue up for instance either before the office is opened in the morning or before it is reopened after the lunch break. One is now interested in:

(i) waiting times for customers from both priority classes to be not too long;

(ii) the initial busy period (starting with k customers in the system) to be not too short.

Small values of k will satisfy (i) whereas large values satisfy (ii).

In [1] the transient waiting times for customers belonging to both priority classes are arrived at for the general model. Using this results on the special arrival pattern above, optimal values for k have been tabulated for various values of p, λ and service intensity.

COMPUTER TIME-SHARING

In [2] a modification of the socalled round-robin priority discipline is treated in an $M/M/1$ queueing model. Each program receives a quantum q of service at a time from a single central processor. If this completes its service requirement, it leaves the system. If not, and there is a new arrival during the service of the quantum, it is given an additional quantum. Otherwise the program joins the end of the queue to wait for its next turn. The model is analysed under the assumption of a constant, nonzero

overhead when the processor swaps one program for another. Obviously, during periods of high arrival rates, this algorithm has the effect of reducing the system's swapping activities. On the other hand, during periods of low arrival rates the discipline is similar to the conventional round-robin, which automatically gives priority to programs with lesser service time requirements.

[2] arrives at expressions for the mean waiting time in queue as a function of the quanta required, and for the mean system cost due to waiting. Numerical comparisons with the conventional round-robin discipline are performed.

A TELEPHONE EXCHANGE

In [6] a telephone exchange is considered handling the calls to and from a minor group of subscribers. Let the latter group belong to the second priority class whereas people from the rest of the world belong to the first class. The Electronic Selector Bar Operator (ESBO) serves customers from both classes. The model one wished to be able to analyse is a modified version of GI/G/M/G/1 with a nonpreemptive priority discipline. However, customers of the higher priority are just allowed to wait a fixed length of time before they hear the busy signal and are lost. This is not the case for customers of lower priority. Secondly, whereas we have a FIFO queueing discipline within the first class, it is RANDOM within the second, i.e. all customers waiting have equal chance of being served next.

Unfortunately, there seems no way of arriving at the stationary waiting time distributions for customers of the two classes by using queueing theory of today. As a first approximation [6] starts out from an $M_1/G/M_2/G/1$ model where Laplace Transforms of

the above distributions are well-known. Furthermore, in this case it is possible to get a look behind the "Laplace curtain".

References

[1] Dalen, G. and Natvig, B. (1980). J. Appl.Prob., 17, 227-234.

[2] Heacox, H. and Purdom, P. (1972), J. Ass. Comp.Mach., 19, 70-91.

[3] Hokstad, P. (1978). INFOR, 16, 158-170.

[4] Jaiswal, N.K (1968). Priority Queues. Academic Press, New York.

[5] Kleinrock, L. (1976). Queueing Systems, Vol.2: Computer Applications. Wiley, New York.

[6] Natvig, B. (1977). On the waiting-time for a priority queueing model. Tech.Rep.No.3, Dept. of Mathematical Statistics, University of Trondheim. Prepared for the Norwegian Telecommunications Administration Research Establishments.

(QUEUEING THEORY)

Bent Natvig
University of Oslo